CLAIMS

What is claimed is:

1. A method comprising:

inputting a vector space;

inputting a probability space; and

generating a similarity space.

2. The method of claim 1 wherein said vector space comprises an eigenspace analysis.

3. The method of claim 1 wherein said vector space comprises a principal component analysis.

4. The method of claim 1 wherein said probability space comprises a transition probability space.

5. The method of claim 1 wherein said probability space comprises a Markov model.

6. The method of claim 1 wherein said vector space represents a reference, wherein said probability space represents a target, and said similarity space represents an indication of similarity between said target and said reference.

7. A machine-readable medium having stored thereon instructions, which when executed performs the method of claim 1.

8. A system comprising a processor coupled to a memory, which when executing a set of instructions performs the method of claim 1.

9. The method of claim 1 further comprising communicating a payment and/or credit.

10. An apparatus for deriving a similarity measure comprising:

means for inputting an eigenspace analysis of a reference;

means for inputting a transition probability model of a target; and

means for operating on said eigenspace analysis and said transition probability model.

11. A machine-readable medium having stored thereon information representing the apparatus of claim 10.

12. An apparatus comprising:

a first block having an input and an output, said first block input capable of receiving a vector space;

a second block having an input and an output, said second block input capable of receiving a probability space; and

a third block having a first input, a second input, and an output, said third block first input coupled to receive said first block output, said third block second input coupled to receive said second block output, and said third block output capable of communication a similarity space.

13. A machine-readable medium having stored thereon information representing the apparatus of claim 12.

14. A method comprising:

generating a similarity metric based upon an eigenspace analysis and an n-gram model.

15. A method comprising:

receiving a profile;

receiving a matrix; and

generating a similarity indication between said profile and said matrix.

16. The method of claim 15 wherein said profile is an eigenspace.

17. The method of claim 15 wherein said matrix is a transition probability matrix.

18. The method of claim 15 wherein said profile is derived from tokens.

19. The method of claim 15 wherein said matrix is derived from tokens.

20. The method of claim 15 wherein said profile is derived from an n-gram analysis of text.

21. A method for generating a similarity space, the method comprising:

combining a vector space with a transition probability space.

22. A method comprising performing a mathematical operation using an eigenspace and transition probability matrix to generate a similarity index.

23. A method for generating a similarity score comprising:

receiving a profile having eigenvalues (w.i);

receiving a transition matrix (a);

generating a left eigenvector (u.i) for each said w.i;

generating a right eigenvector (v.i) for each said w.i;

generating a complex conjugate of said u.i; and

generating said similarity score according to a formula:

similarity score = sum(i, || u.i.conjugate * a * v.i ||^2),

where each term of said summation is said transition matrix premultiplied by said complex conjugate of i-th said left eigenvector, and postmultiplied by i-th

said right eigenvector and wherein norm squared ||.||^2 is a square of magnitude of a complex number resulting from said i-th term in said sum.

24. The method of claim 23 wherein said profile has parameters selected from the group consisting of tuples, tokens, eigenvalues, left eigenvectors, and right eigenvectors.

25. The method of claim 23 wherein said transition matrix is a probability transition matrix.

26. The method of claim 23 wherein said profile represents a reference text and said transition matrix represents a target text, and a lower similarity score indicates that said target text has little or nothing in common with said reference text versus a higher similarity score indicating that there are common tuple-token combinations that appear in said target text that also appear in said reference text.

27. An apparatus for generating a similarity measure comprising:

means for performing a computation on a target input and a profile.

28. The apparatus in claim 27 wherein said computation is substantially linear in order of magnitude with respect to a plurality of target inputs against said profile.

29. The apparatus of claim 27 wherein said target input comprises a transition probability matrix of said target input tokenized.

30. The apparatus of claim 27 wherein said profile comprises an Eigenspace.

31. A means for computing a similarity measure wherein said computing means is substantially O(log(n)) for n target inputs against a pre-computed profile.

32. The means of claim 31 wherein said pre-computed profile is an eigenspace and said one or more target inputs are one or more transition probability matrices.

33. The means of claim 32 wherein said one or more transition probability matrices are derived from one or more sets of shuffled tokens from one or more target inputs.

34. A method comprising:

tokenizing a target input;

generating a transition probability matrix for said tokens;

operating on a profile and said transition probability matrix; and

generating a measure of similarity between said profile and said matrix.

35. The method of claim 34 wherein said profile comprises:

tokenizing a reference input;

generating a transition probability matrix for said reference tokens; and

generating an eigenspace for said transition probability matrix for said reference

tokens.

36. The method of claim 34 wherein said target input is selected from the group consisting of letters, groups of letters, words, phrases, sentences, paragraphs, sections, spaces, punctuation, one or more documents, XML, textual input, HTML, SGML, and sets of text.

37. The method of claim 35 wherein said reference input is selected from the group consisting of letters, groups of letters, words, phrases, sentences, paragraphs, sections, spaces, punctuation, one or more documents, XML, textual input, HTML, SGML, and sets of text.

38. A method for modeling comprising;

using a history window of h tokens to compose a tuple; and

tallying all words that fall within r tokens of said tuple wherein r is between r=1 (which is a Markov n-gram model), and r substantially approaching infinity (which is a word frequency model).

39. The method of claim 38 wherein said r is a step function token transition window of width r.

40. The method of claim 38 wherein said r is a non-step function.

41. The method of claim 40 where said non-step function gives greater weight to nearby tokens and lesser weight to farther away tokens.

42. The method of claim 38 wherein said r is a transition weight function s(i), where 0<=s(i)<=1, for i=1,□ ,r, and normalized so that sum(i=1,□ ,r; s(i)) = 1.

43. A method for generating a similarity measure between a reference profile and a target input by performing an operation on an eigenvalue space representation of said reference profile and a transition probability model of said target input.

44. The method of claim 43 wherein said transition probability model represents a tokenized representation of said target input.

45. The method of claim 44 wherein said tokenized representation is further generated by shuffling of tokens representing said target input.

46. The method of claim 45 wherein a plurality of similarity measures is generated based on said reference profile and one or more said shuffled tokenized representations as said transition probability model.

47. A method for determining a high similarity measure, the method comprising:

(a) pre-generating a fixed set of eigenspace profiles representing known references;

(b) generating a series of tokens representing clauses from a target input;

(c) dividing said series of tokens into two groups, group A and group B;

(d) generating a transition probability model for group A and group B;

(e) generating a similarity measure for group A versus said profiles, and for group B versus said profiles;

(f) retaining group A if it has a similarity measure equal to or higher than group B from (e), otherwise retaining group B; and

(g) define the retained group as said series of tokens and repeat (c) to (g) for a predetermined number of times.

48. The method of claim 47 wherein said (c) dividing said series of tokens into two groups, group A and group B results in group A and group B being substantially the same size.

49. The method of claim 47 wherein said predetermined number of times is based upon a factor selected from the group consisting of a relationship to the number of said tokens representing clauses from said target input, and a predetermined minimum similarity measure.

50. The method of claim 47 wherein said dividing further comprises shuffling said tokens.

51. The method of claim 47 wherein said known references comprises N text blocks.

52. A method comprising:

receiving N text blocks;

building a binary tree representing indexes of said N text blocks;

receiving a T text block;

computing a transition probability matrix for said T text block; and

traversing said binary tree; and

finding a closest matching N text block for said T text block.

53. The method of claim 52 wherein said building further comprises:

concatenating said N text blocks;

computing a profile of said concatenated N text blocks; and

computing partitioning eigenvectors of said N text blocks.

54. A method comprising:

receiving a T text block;

computing a profile for said T text block;

receiving N text blocks;

(a) shuffling randomly said N text blocks;

(b) dividing said shuffled randomly N text blocks into set A and set B;

(c) concatenating the text clocks in set A to form group A;

(d) concatenating the text clocks in set B to form group B;

(e) computing a transition probability matrix for said group A and for said group B;

(f) generating a similarity measure between said T text block and said group A and said group B;

(g) determining if group A or group B has a higher similarity measure;

(h) tallying an additional count for text blocks that are members of group A or group B having said determined higher similarity measure;

(i) repeating (a) through (h) R times;

(j) picking group A or group B with a highest count as a remaining group;

(k) using the remaining group now as said N text blocks; and

(l) repeating (a) through (k) K times.

55. The method of claim 54 wherein group A and group B are substantially a same size.

56. The method of claim 54 wherein R is less than 1025.

57. The method of claim 54 wherein R is determined dynamically.

58. The method of claim 54 wherein K is determined from the group consisting of dynamically, and a fixed count.

59. A method for determining a high similarity measure, the method comprising:

pre-generating one or more eigenspace profiles representing clauses from a known reference;

generating a series of tokens representing clauses from a target input;

(a) setting a counter n=0;

(b) setting counter n=n+1;

(c) dividing said series of tokens into two groups, group A(n) and group B(n);

(d) generating a transition probability model for group A(n) and group B(n);

(e) generating a similarity measure for group A(n) versus said profiles, and for group B(n) versus said profiles;

(f) awarding group A(n) a point if it has a similarity measure equal to or higher than group B(n) from (e), otherwise awarding group B(n) a point;

(g) shuffling said series of tokens representing clauses from said target input in substantially random order and repeat (b) to (g) for a predetermined number of times;

(h) picking those groups having a point and retaining tokens associated with said picked groups and defining said retained tokens as said series of tokens; and

(i) repeating (c) to (h) until said high similarity measure is determined.

60. The method of claim 59 wherein said (c) dividing said series of tokens into two groups, group A and group B results in group A and group B being substantially the same size.

61. The method of claim 59 wherein said predetermined number of times is based upon a factor selected from the group consisting of a relationship to the number of said tokens representing clauses from said target input, and a predetermined minimum similarity measure.

62. The method of claim 59 wherein said method of claim 59 computation is substantially M*log(N), where N represents said clauses from said known reference, and where M represents said clauses from said target input.